Medical case retrieval from a committee of decision trees

Gwénolé Quellec, Mathieu Lamard, Lynda Bekri, Guy Cazuguel, Christian Roux, Béatrice Cochener

Medical case retrieval from a committee of decision trees

Gwénolé Quellec, Mathieu Lamard, Lynda Bekri, Guy Cazuguel, *Member, IEEE,* Christian Roux, *Fellow member, IEEE,* Béatrice Cochener

***Abstract*—A novel content-based information retrieval framework, designed to cover several medical applications, is presented in this paper. The presented framework allows the retrieval of possibly incomplete medical cases consisting of several images together with semantic information. It relies on a committee of decision trees, decision support tools well suited to process this type of information. In our proposed framework, images are characterized by their digital content. It was applied to two heterogeneous medical datasets for computer aided diagnosis: a diabetic retinopathy follow-up dataset (DRD) and a mammography screening dataset (DDSM). Measure of precision among the top five retrieved results of 0.788 ± 0.137 and 0.869 ± 0.161 was obtained on DRD and DDSM, respectively. On DRD for instance, it increases by half the retrieval of single images.**

***Index Terms*—information retrieval, decision trees, CBIR, CAD, medical databases**

I. Introduction

MEDICAL experts base their diagnoses on a mixture of textbook knowledge and experience acquired through real-life clinical cases, hence the growing interest in Case-Based Reasoning (CBR) [1] for computer aided diagnosis systems [2]. CBR assumes that analogous problems have similar solutions: interpreting a new situation involves retrieving similar cases in a case database. Relevance is usually modeled via a similarity measure between a query (a new medical case analyzed by a medical expert) and each case in a reference database. The retrieved cases are then used to help interpreting the new case [1].

CBR was originally designed to process structured cases such as regular feature vectors. However, information required by physicians to diagnose some pathologies are more complex. To diagnose Diabetic Retinopathy (DR) for instance, physicians analyze series of images together with — usually structured — contextual information, such as the patient age, sex and medical history [3], [4]. Consequently, medical CBR systems should be able to manage both symbolic information such as clinical annotations, and numerical information such as images. Some existing systems were designed to manage symbolic information [5]. Some others, relying on Content-Based Image Retrieval (CBIR) [6], [7], were designed to manage digital images [8], [9], [10]. However, there were only few attempts to merge these two approaches. One existing system linearly combines a text based and an image based similarity measure into a common similarity measure [11]; however, this approach does not apply to structured textual information. Another system lets the user restrict a CBIR search to images acquired from the same localization and/or with the same device [12]. More generally, another system lets the user restrict a CBIR search to images whose contextual information match an SQL query specified by the user [13]; however, he/she is assumed to know which queries are relevant: it is likely not the case if such a system is needed for diagnosis aid. As a consequence, we believe heterogeneous information retrieval — i.e. information retrieval based on both clinical descriptors and digital image features — is still an open issue. A novel CBR approach that fuses these two types of information is presented in this paper.

In the proposed framework, heterogeneous attributes (digital images, nominal and continuous variables) have to be aggregated and the value of some of these attributes is possibly unknown. To solve this generalized CBR problem, the use of decision trees (DTs) is proposed [14], [15]. A novel indexing scheme based on DTs is introduced; for improved retrieval efficiency, several DTs can be used. To that purpose, a randomized decision tree learning algorithm is applied so that several DTs can be generated. Finally, a boosting strategy is proposed to handle unbalanced classes [16]. The proposed framework has another advantage: the time required for a user (e.g. a medical expert) to query the reference database can be reduced. A procedure is proposed to update the retrieval list as new attributes are inputted by the user. As a consequence, the user can decide to stop inputting attributes if he/she is satisfied with the results. Moreover, each time he/she inputs an attribute, a second procedure identifies the remaining attributes likely to be the most discriminant; in other words, a fast path towards satisfactory results is suggested.

The paper is organized as follows. Section II presents decision trees and their advantages for heterogeneous information retrieval. Section III explains how images can be included in a decision tree. The proposed decision tree based retrieval framework is presented in section IV. Section V describes the medical datasets used for evaluation. Results are given in section VI and we end with a discussion and conclusions in section VII.

II. Decision trees

A. Description

A decision trees (DT) [14], [15] is a decision support tool relying on a set of rules dividing a population of cases into homogenous groups. Each rule associates a conjunction of tests on some attributes with a group (for instance “if sex=male and

Gwénolé Quellec, Guy Cazuguel and Christian Roux are with INSTITUT TELECOM/TELECOM Bretagne, Dpt ITI, Brest, F-29200 France.

Mathieu Lamard and Béatrice Cochener are with University of Bretagne Occidentale, Brest, F-29200 France.

Gwénolé Quellec, Mathieu Lamard, Lynda Bekri, Guy Cazuguel, Christian Roux and Béatrice Cochener are with Inserm, U650, Brest, F-29200 France.

Lynda Bekri and Béatrice Cochener are with CHU Brest, Service d’Ophtalmologie, Brest, F-29200 France.

age<40 then the case belongs to group 3"). In our case, attributes are either images or contextual information. These rules are organized as a tree; the structure of this tree can be interpreted as follows (see Fig. 1):

- each non-terminal node represents a test on a single attribute (e.g. what is the patient sex ?)
- each edge represents a test outcome (e.g. male)
- each leaf represents a cluster of cases that provided a similar answer to each test (e.g. males younger than 40)

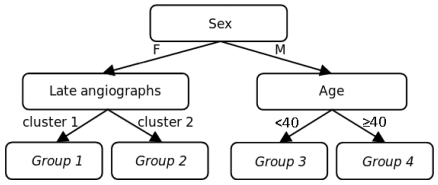

Fig. 1. Toy example of decision tree. Late angiographs are images obtained from one modality (late angiography - see section V-A): in this example, these images are clustered into 2 groups.

DTs were first designed to segment a population of nominal attribute vectors (each test outcome corresponds to an attribute value or group of values). Quinlan [17] extended them to continuous attributes (training cases are grouped by attribute value ranges). More generally, DTs can process any attribute, so long as we can provide a way to cluster cases with respect to that attribute. Since each test is performed on a single attribute, DTs are well suited to process heterogeneous cases.
DTs are generally used as classifiers: an unlabeled case is first associated to a group, and then it is assigned to the most frequent class in that group. In the presented application, we won't use DTs as classifiers; we will use them to define a similarity measure between two cases.

B. Learning

To build a DT in an automatic fashion, we search for the most discriminant attributes and divide the population into homogenous groups according to the value of those attributes (see Fig. 2). This process is supervised and then requires classified cases. In the medical datasets we considered, the disease severity level was used as a class label. Before learning the tree, the dataset has to be divided into three subsets:

- a learning set L, actually used to learn the DT (at the end of the learning process, each case in this set is assigned to the leaves of the tree),
- a validation set V, used to assess the performance of the DT with different parameter settings,
- a test set T, used to assess the final performance of the DT, using the optimal parameter setting

Note that cases assigned to V and T are not used to learn the DT, and T is not used to tune the system at all.

At the beginning of the learning process, the tree consists of a single node containing the whole learning set L. Then each leaf l of the growing tree is recursively divided. In that purpose, the most discriminant attribute f among the population $P \subset L$ assigned to leaf l is searched for. P is then divided into new child nodes, one for each possible answer to the test on f. In the proposed method, the discriminant power of a test is measured by the Shannon entropy gain G obtained when dividing a node v_0 into its child nodes v_n, n=1..N (c4.5 algorithm [14], see equation (1)).

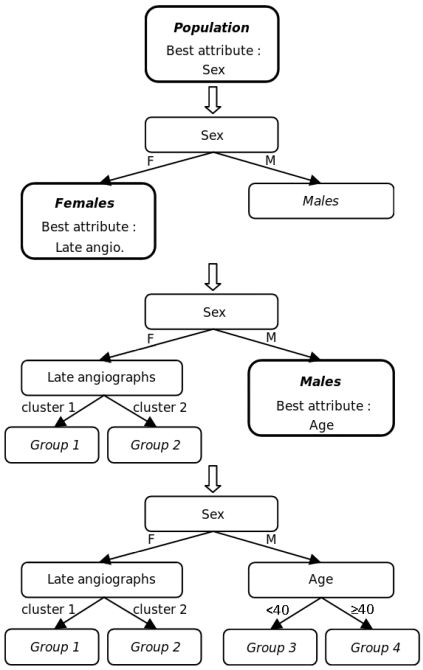

Fig. 2. Illustration of the learning process. At each step, a group of medical cases is divided into subgroups, according to the value of the most discriminant attribute within that group.

$$\begin{cases} G = \left(\sum_{n=1}^{N} I^n\right) - I^0 \\ I^n = -\sum_{c=1}^{C} p_{cn} \log p_{cn},\ n = 0..N \end{cases} \quad (1)$$

where p_{cn} is the percentage of cases assigned to class c in node v_n, $c = 1..C$, I^0 is the entropy in node v_0 (before it is split) and I^n is the entropy in the n^{th} child node v_n, $n = 1..N$. Entropy measures the homogeneity of each node with respect to class label. If no test can improve the entropy enough or if population P is too small, l is not divided.
The learning algorithm can manage missing information: we describe herein the mechanism provided by c4.5 algorithm [14]. Suppose that the value of an attribute f, tested at a node v_0, is missing for a case. Then this case is assigned to each child v_n of v_0 with a weight $w(e_{0n})$, $0 \leq w(e_{0n}) \leq 1$, where e_{0n} denotes the edge from v_0 to v_n. $w(e_{0n})$ is the percentage of cases in v_0, whose value for f is known, assigned to v_n (see Fig. 3). In other words, $w(e_{0n})$ approximates the prior probability for a case in v_0 to belong to v_n. Consequently, at the end of the learning process, each learning case c_i is assigned to each leaf l_j, $j = 1..M$, with a weight w_{ij} such that $\sum_{j=1}^{M} w_{ij} = 1$ (w_{ij}=0 or 1 if each tested attribute is known for c_i, $0 < w_{ij} < 1$ otherwise).

III. Including images in a Decision Tree

To include images in a DT, the principle of Content-Based Image Retrieval (CBIR) is applied [7]. CBIR involves 1) building a feature vector characterizing each image — this feature vector is referred to as signature —, and 2) defining a distance measure between two signatures. In the proposed framework, images are

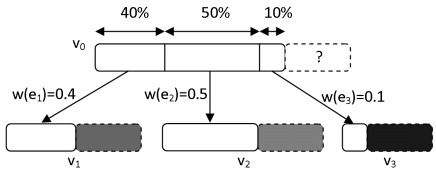

Fig. 3. Managing missing information. If the tested attribute is unavailable for some medical case (?), this case is assigned to all subgroups with a weight equal to the prior probability to be assigned to that subgroup.

characterized by their wavelet transform [18]. Then, measuring the distance between two images comes down to measuring the distance between their signatures. Similarly, when building a DT, we use the distance measure between image signatures to divide a population of images into subgroups. An unsupervised classification algorithm is used to cluster similar image signatures, as described in section III-C. By this process, image signatures can be included in a DT like any other attribute.

A. Image signature

In previous studies on CBIR, we decided to extract signatures for images from their wavelet transform [18]. Using the wavelet transform for database management is convenient: images can be compressed in JPEG-2000 format [19], which relies on the wavelet transform, and their signature can be extracted directly in the compressed domain. Moreover, wavelet-based image signatures have shown their superiority over other image signatures [18]. The proposed signatures model the distribution of the wavelet coefficients in each subband of the wavelet decomposition; as a consequence, a multiscale description of images is obtained. To characterize the distribution of wavelet coefficients in a given subband, Wouwer's work was applied [20]: Wouwer showed that this distribution can be modeled by a generalized Gaussian function (see equation (2)).

$$p(x;\alpha,\beta) = \frac{\beta}{2\alpha\Gamma(\frac{1}{\beta})} e^{-\left(\frac{|x|}{\alpha}\right)^{\beta}} \tag{2}$$

$$\Gamma(z) = \int_0^{\infty} e^{-t} t^{z-1} dt, z > 0 \tag{3}$$

The maximum likelihood estimators $(\hat{\alpha}, \hat{\beta})$ of the wavelet coefficient distribution in each subband are used as a signature [21]. These estimators can be computed directly from JPEG-2000 compressed images, which can be useful when a large number of images have to be processed. Any wavelet basis can be used to decompose images. However, the effectiveness of the extracted signatures largely depends on the choice of this basis. For this reason, we proposed to search for an optimal wavelet basis within the lifting scheme framework [18], which is at the core of the JPEG-2000 compression standard.

B. Distance Measure

Do and Vetterli proposed the use of the Kullback-Leibler (see equation (4)) divergence between wavelet coefficient distributions in each subband to define a distance measure between signatures [21].

$$D(p(X;\theta_1)||p(X;\theta_2)) = \int p(X;\theta_1) \log \frac{p(X;\theta_1)}{p(X;\theta_2)} dx \tag{4}$$

Kullback-Leibler divergence is not symmetric, which is a requirement of clustering algorithms. A symmetric version of the divergence, Ds, is used instead (see equation (5)).

$$\frac{Ds(p(X;\theta_1)||p(X;\theta_2)) = }{\frac{(D(p(X;\theta_1)||p(X;\theta_2)) + D(p(X;\theta_2)||p(X;\theta_1)))}{2}} \tag{5}$$

By injecting equation (2) in (5), we obtain the expression of the distance measure between two wavelet coefficient distributions (see equation (6), the expression in the asymmetrical case is given in [21]).

$$\begin{aligned} &Ds(p(X;\alpha_1,\beta_1)||p(X;\alpha_2,\beta_2)) = \\ &\left(\frac{\alpha_1}{\alpha_2}\right)^{\beta_2} \frac{\Gamma(\frac{\beta_2+1}{\beta_1})}{\Gamma(\frac{1}{\beta_1})} + \left(\frac{\alpha_2}{\alpha_1}\right)^{\beta_1} \frac{\Gamma(\frac{\beta_1+1}{\beta_2})}{\Gamma(\frac{1}{\beta_2})} - \frac{1}{\beta_1} - \frac{1}{\beta_2} \end{aligned} \tag{6}$$

Finally, the distance between two images is a weighted sum of these symmetric divergences over the subbands [18]. The ability to select a weight vector and a wavelet basis makes this image representation suitable for specialized medical datasets.

C. Signature Clustering

Thanks to the image signatures and the associated distance measure above, a population of cases can be divided into subgroups using an unsupervised classification algorithm, provided that a custom distance measure can be specified. Because it is simple and fast, the Fuzzy C-Means algorithm (FCM) [22] was used for this purpose; the Euclidian distance was replaced in FCM by the proposed distance between signatures. Finding the right number of clusters is generally a difficult problem. However, when the data is labeled, mutual information between cluster and class labels can be used to determine the optimal number of clusters $\hat{K}$ [23] (see equation (7)).

$$\hat{K} = \underset{K}{\operatorname{argmax}} \sum_{c=1}^{C} \sum_{k=1}^{K} p(c,k) \log_{C+K} \frac{p(c,k)}{p(c)p(k)} \tag{7}$$

where $c = 1..C$ are the class labels, $p(c,k)$ is the joint probability distribution function of the class and cluster labels, $p(c)$ and $p(k)$ are the marginal probability distribution functions.

IV. Decision Tree Based Retrieval

A. Objective

Let c_q be a case placed as a query by a user. The objective of the proposed framework is to retrieve the R most similar cases in a reference database. For diabetic retinopathy follow-up, the number of cases retrieved by the system is set to $R = 5$, at ophthalmologist's request; they consider this number sufficient for time reasons and in view of the results provided by the system. Consequently, the satisfaction of the user's needs is assessed by the precision at R, denoted π_R, defined as the percentage of cases relevant for c_q among the topmost R results.

B. Single Tree Based Indexing

To find the R most similar cases, we need to compute a similarity measure between c_q and each case c_i in the reference database. To that purpose, we propose to compare their assignment weights to each leaf l_j: w_{qj} and w_{ij}, $j = 1..M$. These weights have been computed for each learning case (subset L of

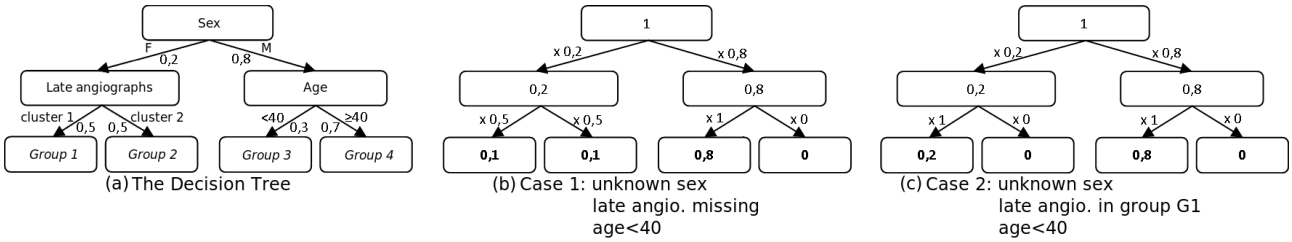

Fig. 4. Illustration of the retrieval process. In (b), where attribute 'late angio.' is missing, the leaves for Groups 1 to 3 will be browsed, whereas for (c), where this attribute is available, only the leaves for Groups 1 and 3 will be browsed.

the reference database) while building the tree (see section II-B). They can be computed a posteriori for each remaining case in the reference database — in particular those added after the learning phase — and for the query c_q. In that purpose, the weight $w(e)$ of each edge e in the tree is stored (see section II-B). The retrieval system is illustrated on an example in Fig. 4.

A similarity measure S_{ab} between two cases c_a and c_b is defined in equation (8); S_{ab} relies on the assignment weight $(w_{aj})_{j=1..M}$ (resp. $(w_{bj})_{j=1..M}$) of c_a (resp. c_b) to each to each leaf l_j, $j = 1..M$.

$$S_{ab} = \sum_{j=1}^{M} w_{aj} w_{bj} \tag{8}$$

This similarity measure, the scalar product of $(w_{aj})_{j=1..M}$ and $(w_{bj})_{j=1..M}$, maps to [0;1]. It is maximal when a and b are completely assigned to the same leaf. It is minimal when there is no leaf to which both cases are at least partially assigned.
The similarity measure between the query c_q and each case c_i in the reference database can be computed very quickly. It does not require browsing the entire reference database:

- For each leaf l_j in the DT, a list L_j containing every case c_i such that $w_{ij} \neq 0$ is built during the learning process. These lists are updated each time a new case is included in the reference database.
- At the beginning of the retrieval process, each similarity measure S_{qi} is set to 0.
- For each leaf l_j such that $w_{qj} \neq 0$, L_j is browsed: for each case $c_i \in L_j$, S_{qi} is increased by $w_{qj} w_{ij}$.

C. Multiple Tree Based Indexing

Because of the hierarchical architecture of the system above, some attributes might be given too much weight. In the example of Fig. 4 for instance, a male and a female both aged 30 would be regarded as completely dissimilar, because of their different sex, whereas age might play a significant role. To solve this problem, we propose a retrieval system relying not only on one DT but rather on several (say τ) DTs. Retrieving similar cases from a single DT or from several DTs can be done similarly: instead of computing the scalar product between the assignment weights to the leaves of one DT, we simply compute the scalar product between the assignment weights to the leaves of each of these τ DTs. The expression of the new similarity measure S'_{ab} is given in equation 9.

$$S'_{ab} = \sum_{t=1}^{\tau} \sum_{j=1}^{M_t} w_{atj} w_{btj} \tag{9}$$

where w_{atj} is the assignment weight of case c_a to the j^{th} leaf of the t^{th} tree and M_t is the number of leaves in the t^{th} tree.
Several methods have been proposed in the literature to generate sets, or committees, of DTs: Random Forests [24] or randomized c4.5 [25] for instance. They usually perform better as classifiers than single DTs. To generate DT committees, the learning algorithm is randomized as follows: to decide which test should be selected for dividing a tree node, the k most discriminant attributes, according to the entropy measure (see equation (1)), are identified one of them is picked randomly with uniform probability.

D. Retrieval System Boosting

When applied to unbalanced datasets, DTs tend to be biased towards the largest classes [26]. If DTs are used as classifiers, this problem can be alleviated thanks to boosting [16]. Boosting algorithms typically build a DT committee in iterations, by incrementally adding weak classifiers (i.e. with a predictive accuracy at least better than chance) to a final strong classifier. At each iteration k, a weak classifier h_k is learnt from the learning set with respect to a distribution (learning cases are assigned more or less weight); the weight distribution is initially uniform. The weak classifier is then added to the final strong classifier and the learning cases are reweighted: misclassified cases gain weight and correctly classified cases lose weight. We followed the example of Adaboost [16], the most popular boosting algorithm, to define a boosting strategy for our retrieval system. In our application, h_k denotes a set of DTs used as a "weak retriever" (see section IV-C). At each iteration k, the weight $d_k(c_i)$ of case c_i is updated as follows:

1) the weighted average retrieval error ϵ_k of h_k is computed: $\epsilon_k = 1 - \sum_i d_k(c_i) \pi_R^k(c_i)$
2) the weight α_k of h_k is updated: $\alpha_k = \frac{1}{2} \ln \frac{1-\epsilon_k}{\epsilon_k}$
3) a variable $\gamma_k(c_i)$ indicating whether $d_k(c_i)$ should be increased ($\gamma_k(c_i) > 0$) or decreased ($\gamma_k(c_i) < 0$) is computed: $\gamma_k(c_i) = 1 - 2\pi_R^k(c_i)$
4) $d_k(c_i)$ is updated according to α_k and $\gamma_k(c_i)$: $d_{k+1}(c_i) \propto d_k(c_i) e^{\alpha_k \gamma_k(c_i)}$

The final "strong retriever" H is thus a set of DT sets h_k weighted by α_k. Equivalently, H is a DT set in which each tree t in h_k is assigned a weight $\alpha_t = \alpha_k$ in H. Consequently, the final similarity measure becomes S''_{ab}, given in equation (10).

$$S''_{ab} = \sum_{t=1}^{\tau} \sum_{j=1}^{M_t} \alpha_t w_{atj} w_{btj} \tag{10}$$

TABLE I
STRUCTURED CONTEXTUAL INFORMATION FOR DIABETIC RETINOPATHY PATIENTS

	attributes	possible values
general clinical context	family clinical context	diabetes, glaucoma, blindness, misc.
	medical clinical context	arterial hypertension, dyslipidemia, protenuria, renal dialysis, allergy, misc.
	surgical clinical context	cardiovascular, pancreas transplant, renal transplant, misc.
	ophthalmologic clinical context	cataract, myopia, AMD, glaucoma, unclear medium, cataract surgery, glaucoma surgery, misc.
circumstances, examination and diabetes context	diabetes type	none, type I, type II
	diabetes duration	$<$ 1 year, 1 to 5 years, 5 to 10 years, $>$ 10 years
	diabetes stability	good, bad, fast modifications, glycosylated hemoglobin
	treatments	insulin injection, insulin pump, anti-diabetic drug + insulin, anti-diabetic drug, pancreas transplant
eye symptoms reported before the angiography test	ophthalmologically symptomatic	none, systematic ophthalmologic screening - known diabetes, recently diagnosed diabetes by check-up, diabetic diseases other than ophthalmic ones
	ophthalmologically asymptomatic	none, infection, unilateral decreased visual acuity (DVA), bilateral DVA, Neovascular glaucoma, intra-retinal hemorrhage, retinal detachment, misc.
	maculopathy	focal edema, diffuse edema, none, ischemic

V. APPLICATION TO TWO MEDICAL DATASETS

The proposed framework has been applied to two heterogeneous medical datasets. The first dataset (DRD) is being built at the LaTIM laboratory (Inserm U650), in collaboration with ophthalmologists from Brest University Hospital. The second one (DDSM) is a well-known public access dataset [27].

A. Diabetic Retinopathy Dataset (DRD)

Diabetic retinopathy is damage to the retina caused by complications of diabetes, which can eventually lead to blindness. The diabetic retinopathy dataset contains retinal images of diabetic patients, with associated anonymized information on the pathology. The dataset consists of 86 patient files containing 1399 photographs altogether. Patients have been recruited at Brest University Hospital since June 2003 and images were acquired by experts using a Topcon Retinal Digital Camera (TRC-50IA) connected to a computer. Images have a definition of 1280 pixels/line for 1008 lines/image. They are lossless compressed images. An image series is given in Fig. 5. The contextual information available is the patients' age and sex and structured medical information (see table I). If patients records were comprehensive, they would consist of 10 images per eye (see Fig. 5) and of 13 contextual attributes. However, in our dataset, 11.9% of images and 39.7% of contextual attribute values are missing. The disease severity level, according to ICDRS classification [3], was assessed by one three-year experienced expert for each patient. The distribution of the disease severity among the above-mentioned 86 patients is given in table II.

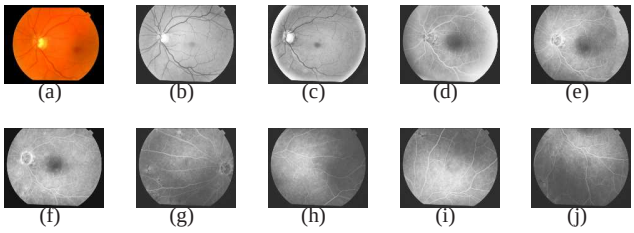

Fig. 5. Photograph sequence of a patient's eye. Photographs (a), (b) and (c) were obtained with different color filters. Photographs (d) to (j) constitute a temporal angiographic series: a contrast product is injected and photographs are taken at different stages (early (d), intermediate (e), (g)-(j) and late (f)). (g)-(j) are images from the periphery of the retina.

TABLE II
PATIENT DISEASE SEVERITY DISTRIBUTION

dataset	disease severity	number of patients
DRD	no apparent diabetic retinopathy	7
	mild non-proliferative	13
	moderate non-proliferative	25
	severe non-proliferative	15
	proliferative	12
	treated/non active diabetic retinopathy	14
DDSM	normal	695
	benign	669
	cancer	913

B. Digital Database for Screening Mammography (DDSM)

The DDSM project [27], involving the Massachusetts General Hospital, the University of South Florida and the Sandia National laboratories, has built a mammographic image database for research on breast cancer screening. It consists of 2277 patient files. Each one includes two images of each breast, along with some associated patient information (age at time of study, subtlety rating for abnormalities, American College of Radiology breast density rating and keyword description of abnormalities) and image information (scanner, spatial resolution, etc.). The following contextual attributes were included in the system:

- age at time of study
- breast density rating
- digitizer

The remaining attributes were not used, either because they are regarded as useless (date of study, date digitized, etc.) or because they require advanced expert interaction (the description of the lesions visible in images). Images have a varying definition, of about 2000 pixels/line for 5000 lines/image. Each patient file has been graded by a physician. Patients are then classified in three groups: normal, benign and cancer. The distribution of grades among the patients is given in table II.

C. Attributes of a medical case

In those datasets, each patient file consists of a mixture of digital images and contextual information. Contextual attributes (13 in DRD, 3 in DDSM) do not require advanced preprocessing: textual attributes (such as "treatments" in DRD) were translated into codes and processed as nominal attributes; numerical contextual attributes (such as "breast density rating" in DDSM) did not require any preprocessing at all. Images, on the

other hand, require advanced preprocessing: numerical attributes were extracted to characterize them. A usual solution to extract numerical attributes from images is to segment these images and extract domain specific information. However, this approach requires expert knowledge, so it is not sufficient for generic database management. Another solution was proposed: images were indexed by their digital content (i.e. a generic signature was extracted — see section III-A). An image signature was computed for each type of images (10 for DRD, 4 for DDSM). To push performance further, it is possible to include domain-specific knowledge in the proposed framework, in addition to the generic signatures. In DRD, for instance, we counted the number of microaneurysms (the most frequent lesion of diabetic retinopathy) detected by the algorithm described in [28]. We are not mammography experts, so we did not include any expert knowledge in the system for DDSM.

D. Retrieval system calibration

To maximize the retrieval precision of the proposed system, the following parameters had to be tuned:

- the wavelet basis used to decompose images and the weight vector used in the distance measure between signatures (p^0),
- the number p^1 of generated trees in each weak retriever h_k,
- the random parameter $p^2 = k$ (the number of attributes among which the tested attribute is selected at each node),
- the FCM parameter p^3 (the fuzzy coefficient [22])

System performance was assessed by 10-fold cross-validation. Let N be the number of cases in a reference database (N=86 for DRD, N=2277 for DDSM). For each fold, each reference database was divided randomly into two sets: T (the test set — $0.1N$ cases) and $\bar{T}$ ($0.9N$ cases). The wavelet basis and the weights (p^0) were trained on $\bar{T}$ for each "image attribute" [18]. For each element in the product space $P^1 \times P^2 \times P^3$, a DT committee was built using $\bar{T}$. Each tree in this committee was learnt using a learning set L ($0.8N$ cases) selected at random in $\bar{T}$; $V = \bar{T} \setminus L$ was used for validation. The precision at R of this committee for a case c_i in V (i.e. $\pi_R(c_i)$) was computed over the DTs learnt when c_i was in V. The score of this committee is the average $\pi_R(c_i)$, $c_i \in \bar{T}$. The DT committee of maximal performance for the current fold, according to $\bar{T}$, was then assessed using the test set T.

The search for the best element in the product space $P^1 \times P^2 \times P^3$ has been sped up by a genetic algorithm [29].

E. Robustness to information incompleteness

Robustness to information incompleteness has been assessed as follows:

1) For each case c_i in the test set T, 100 new cases have been generated as described in 2).
2) Let n_i be the number of attributes inputted for c_i, each new case has been obtained by removing a number of attribute values randomly selected in $\{0, 1, ..., n_i\}$.
3) Robustness to information incompleteness is visually assessed by plotting the average precision at five with respect to the number of available attributes, using the cases generated as described in 1)-2).

F. Baseline heterogeneous information retrieval method

To evaluate the contribution of DTs for heterogeneous and incomplete case retrieval, the proposed approach has been compared to a weighted sum of heterogeneous distance functions, managing missing values [30]. This method was used as a reference for being the natural generalization of CBR. We extended it to cases containing images thanks to the distance measure between image signatures defined in section III-B.

VI. Results

A precision at R=5 of **0.788±0.137** (resp. **0.869±0.161**) was measured on DRD (resp. DDSM) using the process described in section V-D. It means that, on average, approximately four cases among the five cases retrieved for a query are relevant. The best set of parameter values, obtained for each dataset by the process described in section V-D, is given in table III. Because of the limited number of images per class in DRD (see table II), retrieval performance necessarily drops with increasing Rs: a precision at R=10 of 0.681±0.133 and a precision at R=20 of 0.571±0.129 were obtained on this dataset. Retrieval performance also decreases on DDSM: a precision at R=10 of 0.819±0.157 and a precision at R=20 of 0.756±0.163 were obtained on this dataset. To bring out the discrimination ability of each attribute, we report in table IV the precision at 5 of a retrieval system that simply finds the 5 most similar cases with respect to that attribute. More generally, to estimate the contribution of numerical (image series signatures) and contextual information, DT sets were learnt using numerical or contextual information alone. On DRD, retrieval precision based on all attributes is significantly higher than retrieval precision based on numerical attributes alone, at the 90% confidence level (but not on DDSM). To evaluate the contribution of boosting, the average precision at five over each class, with or without boosting, is given in table V. Robustness to information incompleteness is assessed in Fig. 6. Whereas a precision at five of 0.788±0.137 (resp. 0.869±0.161) was obtained on DRD (resp. DDSM) with the proposed approach, a precision at five of 0.531±0.185 (resp. 0.709±0.251) was obtained on DRD (resp. DDSM) with a usual CBIR approach (i.e. each case simply consists of one image), using the same image signatures; a comparison with other image signatures is provided in [18]. The precision at five obtained with the baseline heterogeneous information retrieval method described in section V-F is 0.553±0.178 on DRD and 0.739±0.182 on DDSM.

Finally, the average computation time required to retrieve the 5 most similar cases, using the settings of table III, is given in table VI. Clearly, most of the time is spent while image signatures are computed. Once the distances between images have been computed, the learning process in itself only takes 0.8 seconds (resp. 80 seconds) per DT for DRD (resp. DDSM). Experiments were performed using an AMD Athlon 64-bit based computer running at 2 GHz.

VII. Discussion and conclusions

A novel medical information retrieval framework has been presented in this paper: it supports queries consisting of image series with contextual information. The framework uses decision

TABLE III
OPTIMAL PARAMETER SETTINGS

parameter	DRD	DDSM
p^1 (nr of trees in h_k)	5	2
p^2 (random parameter)	3	1
p^3 (FCM parameter)	1.5	2.75
total nr of trees	40	16

TABLE IV
INFLUENCE OF EACH ATTRIBUTE ON RETRIEVAL PRECISION

DRD		DDSM	
age	0.321±0.172	age	0.438±0.204
sexe	0.314±0.190	breast density	0.352±0.201
familial clinical context	0.298±0.180	digitizer	0.294±0.198
medical clinical context	0.328±0.195		
surgical clinical context	0.256±0.184		
ophthalmic clinical context	0.353±0.167		
diabetes type	0.319±0.160		
diabetes duration	0.363±0.167		
diabetes stability	0.314±0.186		
treatments	0.302±0.185		
symptoms ophth. symptomatic	0.363±0.164		
symptoms ophth. asymptomatic	0.328±0.179		
maculopathy	0.353±0.159		
contextual	0.463±0.165	contextual	0.451±0.201
number of microaneurysms	0.353±0.187	LCC view	0.682±0.179
green filtered photographs	0.542±0.156	LMLO view	0.685±0.178
blue filtered photographs	0.370±0.178	RCC view	0.694±0.176
red filtered photographs	0.389±0.168	RMLO view	0.691±0.178
early angiographs	0.579±0.149		
interm. angiographs (center)	0.498±0.164		
late angiographs	0.560±0.150		
interm. angiographs (nasal)	0.507±0.165		
interm. angiographs (temporal)	0.474±0.166		
interm. angiographs (upper)	0.449±0.163		
interm. angiographs (lower)	0.519±0.154		
numerical	0.702±0.142	numerical	0.802±0.172
all	0.788±0.137	all	0.869±0.161

TABLE V
PRECISION AT FIVE FOR EACH CLASS

dataset	DRD		DDSM	
class	no boosting	boosting	no boosting	boosting
1	0.307±0.178	0.597±0.150	0.862±0.173	0.873±0.156
2	0.705±0.166	0.802±0.146	0.818±0.183	0.872±0.164
3	0.870±0.096	0.835±0.092	0.839±0.180	0.864±0.162
4	0.719±0.163	0.793±0.156		
5	0.677±0.170	0.758±0.166		
6	0.847±0.128	0.809±0.133		
entire set	0.742±0.157	0.788±0.137	0.840±0.179	0.869±0.161

TABLE VI
COMPUTATION TIME

dataset	DRD	DDSM
wavelet transform (for 1 image)	0.22 s	1.99 s
estimating $(\hat{\alpha}, \hat{\beta})$ (for 1 image)	4.35 s	33.90 s
computing the distance between signatures (for 1 image modality)	0.0335 s	1.14 s
browsing the trees and ranking the cases	0.067 s	0.0032 ms
average total time	17.24 s	99.50 s

trees (DTs) to combine heterogeneous information (in particular, a way to include image signatures in a DT was proposed), handle missing values and avoid over fitting. The latter property, reinforced by boosting, makes this framework well suited to process both large datasets such as DDSM and small datasets such as DRD.

The precision at five obtained for DRD (0.788±0.137) is particularly interesting, in view of the few cases available, the large number of missing values and the number of classes

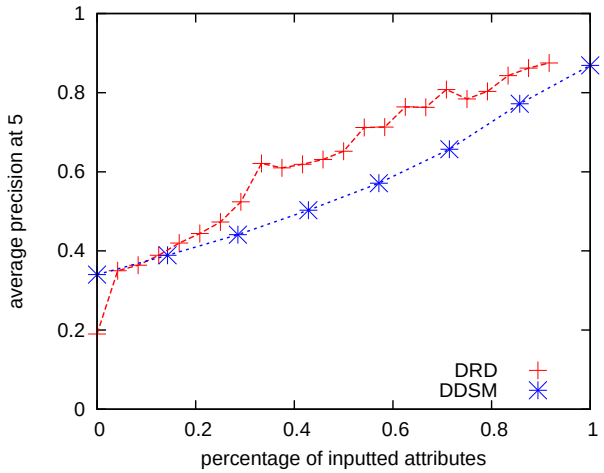

Fig. 6. Robustness to information incompleteness. Good retrieval performance can be obtained even if many attributes are unavailable: on DRD, a precision at five of 60% (resp. 70%) can be reached even if less than 40% (resp. 55%) of the attributes are available.

taken into account (6). On this dataset, the proposed framework outperforms the retrieval of single images by a factor of 1.48 in precision. This stands to reason since an image alone is generally not sufficient for experts to correctly diagnose the disease severity level of a patient. Nevertheless, as table IV shows, images are discriminant attributes, red free photographs (Fig. 5 (b)) and angiographs (Fig. 5 (d)-(j)) in particular. These two image modalities are indeed the most useful for physicians to follow up diabetic retinopathy. On DDSM, the superiority of images over nominal attributes is even more obvious, as illustrated in table IV. This table also shows that using images series without contextual information, instead of single images, increases by itself the average precision at five by a factor of 1.32 on DRD. Adding contextual information increases precision further. Besides, this non-linear retrieval method is 1.43 times more precise than a linear combination of heterogeneous distances on DRD. The improvement brought by heterogeneous information retrieval is more moderate for DDSM (0.869±0.161 as opposed to 0.709±0.251). Performance increase can be explained by the combination of evidence from four images instead of one and by a fine segmentation of the feature space into homogeneous groups provided by DTs, which helps us better separate the classes.

Boosting does not lead to a significant increase of precision over the entire dataset, however it increases precision for rare classes: in DRD, precision significantly increases for class 1, the rarest class, at the 90% confidence level (see table V).

The proposed retrieval system is fast: most of the computation time is spent during the image processing steps. Moreover, it is not necessary to process every image. The first reason is that the retrieval system only needs to characterize attributes tested at nodes browsed by the query case; as a consequence, certain images do not need to be processed. The second reason is that sufficient precision can be reached before every attribute has been inputted by the user. Provided that the retrieval list is updated each time an attribute is updated, the user can stop formulating his/her query when he/she is satisfied with the results. On DRD for instance, a precision at five of 60% can be reached by inputting less than 40% of the attributes (see Fig. 6): with this

precision, the majority of the retrieved cases (3 out of 5) belong to the correct class.

Another interest of the proposed framework is its generality: any multimedia database may be processed so long as a process to cluster cases is provided for each new modality (sound, video, etc). This paper reports promising results about the use of data mining techniques to combine numerical and contextual information in a medical retrieval framework; we are now focusing on alternative data mining algorithms to improve performance.

REFERENCES

[1] A. Aamodt, "Case-based reasoning: Foundational issues, methodological variations, and system approaches," *AI Commun*, vol. 7, no. 1, pp. 39–59, March 1994.

[2] I. Bichindaritz and C. Marling, "Case-based reasoning in the health sciences: What's next?" *Artif Intell Med*, vol. 36, no. 2, pp. 127–135, January 2006.

[3] C. Wilkinson, F. Ferris, R. Klein, and al., "Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales," *Ophthalmology*, vol. 110, no. 9, pp. 1677–1682, September 2003.

[4] M. D. Davis, M. R. Fisher, R. E. Gangnon, F. Barton, L. M. Aiello, E. Y. Chew, F. L. Ferris, and G. L. Knatterud, "Risk factors for high-risk proliferative diabetic retinopathy and severe visual loss: Early treatment diabetic retinopathy study report 18," *Invest Ophthalmol Vis Sci*, vol. 39, no. 2, pp. 233–252, February 1998.

[5] J.-M. Cauvin, C. L. Guillou, B. Solaiman, M. Robaszkiewicz, P. L. Beux, and C. Roux, "Computer-assisted diagnosis system in digestive endoscopy," *IEEE Trans Inform Technol Biomed*, vol. 7, no. 4, pp. 256–262, December 2003.

[6] C. Nastar, "Indexation d'images par le contenu: un état de l'art," in *CORESA'97*, March 1997.

[7] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans Pattern Anal Mach Intell*, vol. 22, no. 12, pp. 1349–1380, December 2000.

[8] H. Müller, N. Michoux, D. Bandon, and A. Geissbuhler, "A review of content-based image retrieval systems in medical applications - clinical benefits and future directions," *Int J Med Inform*, vol. 73, no. 1, pp. 1–23, February 2004.

[9] G. D. Tourassi, R. Vargas-Voracek, D. M. Catarious, and C. E. Floyd, "Computer-assisted detection of mammographic masses: A template matching scheme based on mutual information," *Med Phys*, vol. 30, no. 8, pp. 2123–2130, 2003.

[10] H. Alto, R. M. Rangayyan, and J. E. L. Desautels, "Content-based retrieval and analysis of mammographic masses," *J Electron Imag*, vol. 14, no. 2, p. 023016, 2005.

[11] H. Shao, W.-C. Cui, and H. Zhao, "Medical image retrieval based on visual contents and text information," in *IEEE SMC' 04*, vol. 1, October 2004, pp. 1098–1103.

[12] C. L. Bozec, E. Zapletal, M. Jaulent, D. Heudes, and P. Degoulet, "Towards content-based image retrieval in a HIS-integrated PACS," in *AMIA' 00*, 2000, pp. 477–481.

[13] S. Antani, L. R. Long, and G. R. Thoma, "A biomedical information system for combined content-based retrieval of spine X-ray images and associated text information," in *ICVGIP' 02*, 2002, pp. 242–247.

[14] J. R. Quinlan, *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, 1993.

[15] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, "Classification and regression trees," Wadsworth, Belmont, Ca., 1984.

[16] Y. Freund and R. Schapire, "Experiments with a new boosting algorithm," in *ICML' 96*, July 1996, pp. 148–156.

[17] J. R. Quinlan, "Learning with continuous classes," in *5th Aust. Joint Conf. on Artificial Intelligence*, 1992, pp. 343–348.

[18] G. Quellec, M. Lamard, G. Cazuguel, B. Cochener, and C. Roux, "Wavelet optimization for content-based image retrieval in medical databases," *Med Image Anal*, vol. 14, no. 2, pp. 227–241, April 2010.

[19] D. Taubman and M. Marcellin, *JPEG2000: Image Compression Fundamentals, Standards and Practice (The International Series in Engineering and Computer Science)*. Kluwer Academic Publishers, 2001.

[20] G. van de Wouwer, P. Scheunders, and D. van Dyck, "Statistical texture characterization from discrete wavelet representations," *IEEE Trans Image Process*, vol. 8, no. 4, pp. 592–598, April 1999.

[21] M. N. Do and M. Vetterli, "Wavelet-based texture retrieval using generalized gaussian density and kullback-leibler distance," *IEEE Trans Image Process*, vol. 11, no. 2, pp. 146–158, February 2002.

[22] J. C. Bezdek, "Fuzzy mathemathics in pattern classification," Ph.D. dissertation, Applied Math. Center, Cornell University, Ithaca, 1973.

[23] A. Strehl, "Relationship-based clustering and cluster ensembles for high-dimensional data mining," 2002, doctoral dissertation [electronic resource], The University of Texas at Austin.

[24] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, October 2001.

[25] T. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization," *Machine Learning*, vol. 40, no. 2, pp. 139–157, August 2000.

[26] H. Guo and H. L. Viktor, "Learning from imbalanced data sets with boosting and data generation: the databoost-im approach," *SIGKDD Explor Newsl*, vol. 6, no. 1, pp. 30–39, June 2004.

[27] M. Heath, K. W. Bowyer, and D. K. et al., "Current status of the digital database for screening mammography," in *Digital Mammography, Kluwer Academic Publishers*, 1998, pp. 457–460.

[28] G. Quellec, M. Lamard, P. M. Josselin, G. Cazuguel, B. Cochener, and C. Roux, "Optimal wavelet transform for the detection of microaneurysms in retina photographs," *IEEE Trans Med Imaging*, vol. 27, no. 9, pp. 1230–1241, September 2008.

[29] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*. Kluwer Academic Publishers, Boston, MA, 1989.

[30] D. R. Wilson and T. R. Martinez, "Improved heterogeneous distance functions," *J Artif Intell Res*, vol. 6, no. 1, pp. 1–34, 1997.